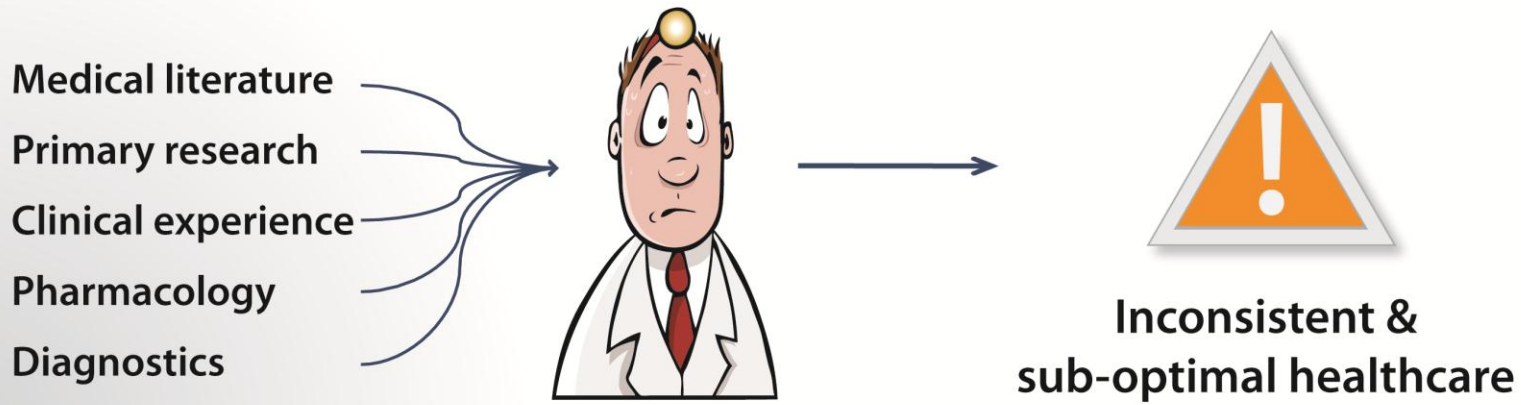
SHARING AND CONNECTING KNOWLEDGE, RATHER THAN (only) DATA

3 October, 2013

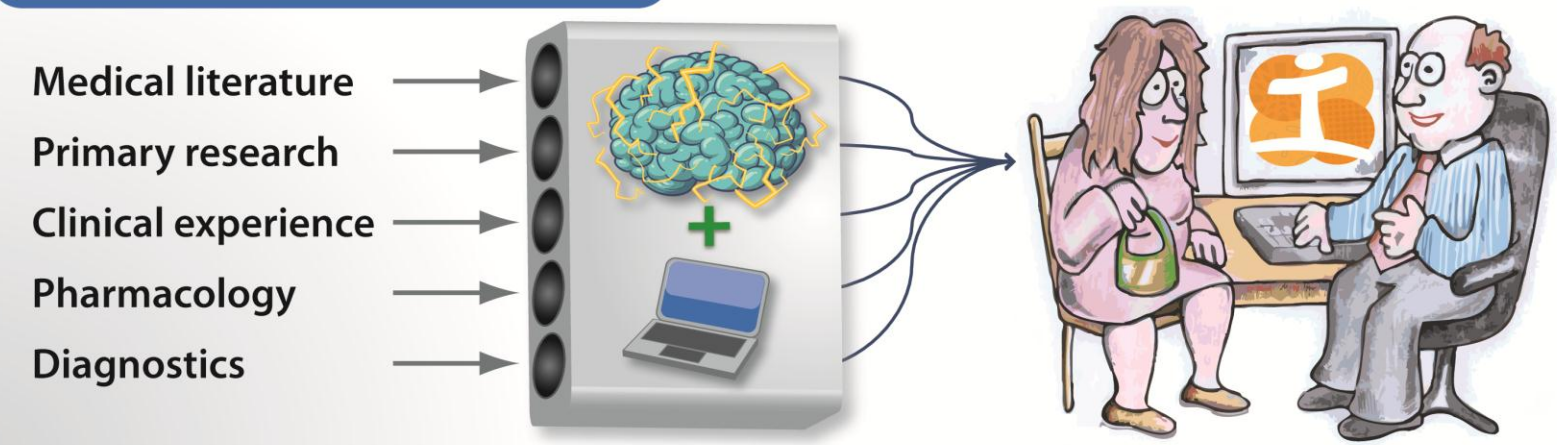
TBC Conference

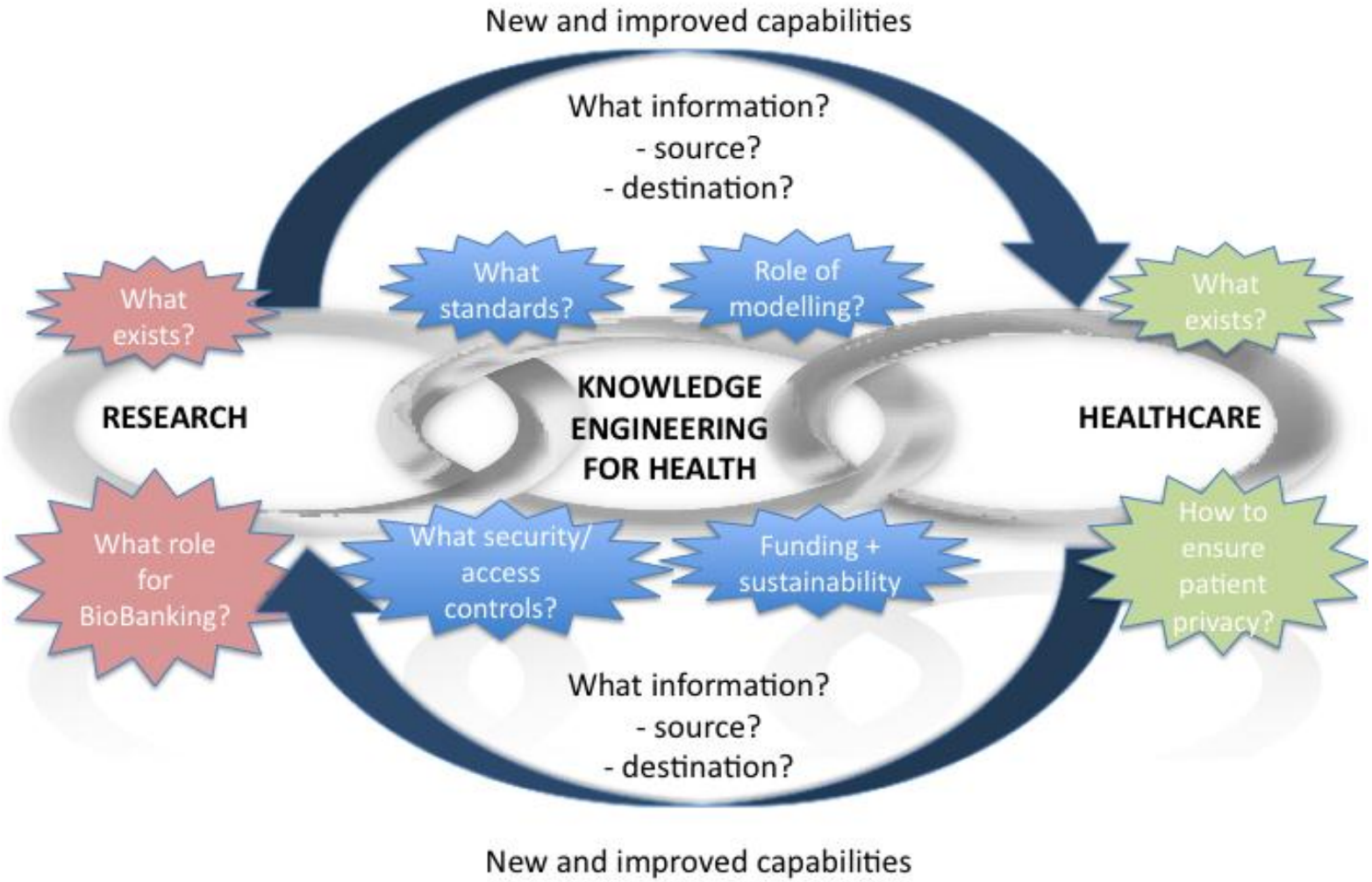
Anthony Brookes: University of Leicester, UK

Today's Healthcare



Tomorrow's Healthcare




New and improved capabilities
What information?
- source?
- destination?
What exists?
What standards?
Role of modelling?
What exists?
RESEARCH
KNOWLEDGE ENGINEERING FOR HEALTH
HEALTHCARE
What role for BioBanking?
What security/ access controls?
Funding + sustainability
How to ensure patient privacy?
What information?
- source?
- destination?
New and improved capabilities

WHY share data?

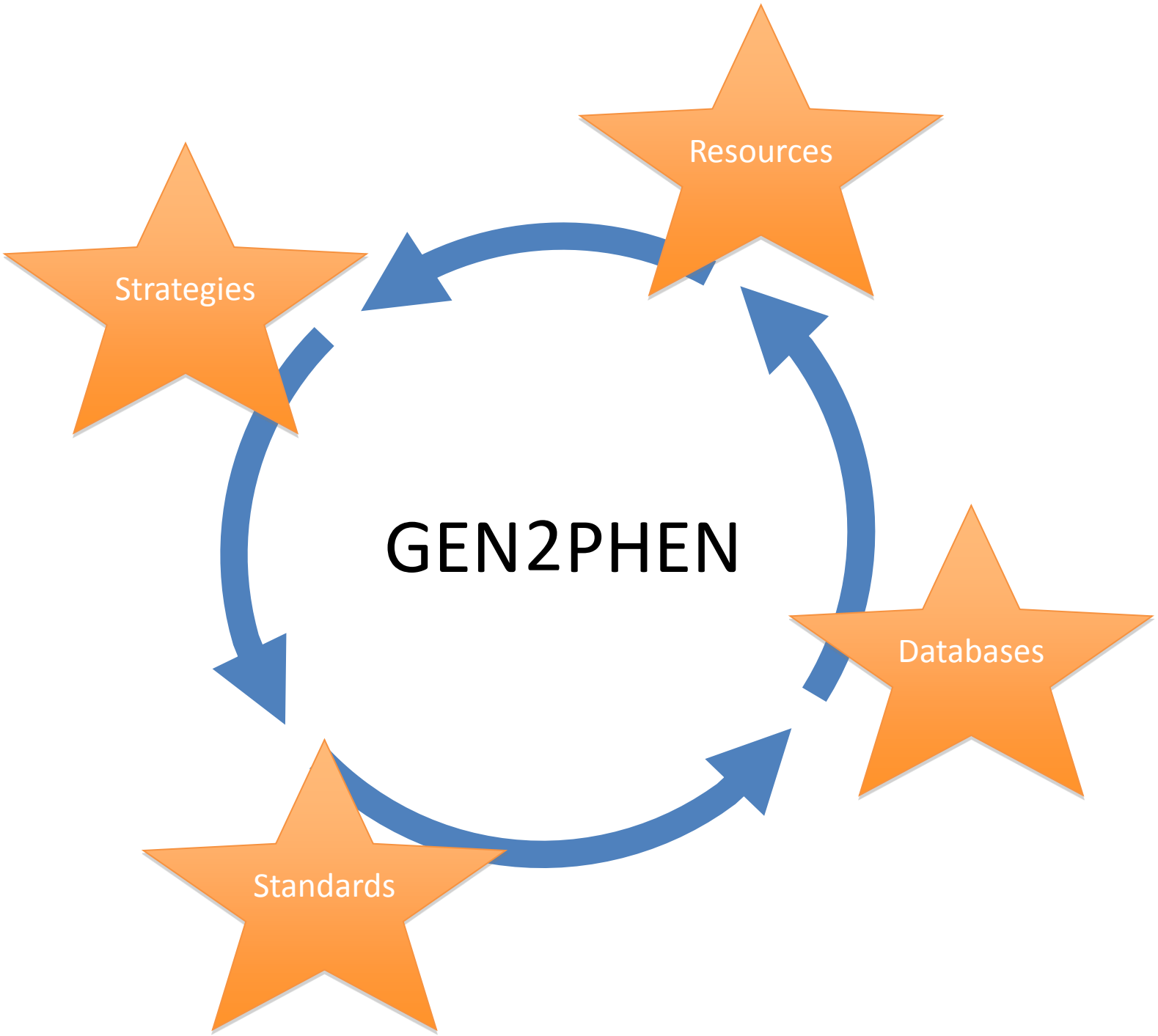
- Enables checking & validation of claims
- Enables meta-analysis
- Enables new uses
- Enables data integration
- . . . .
- Panton principles: “For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open”

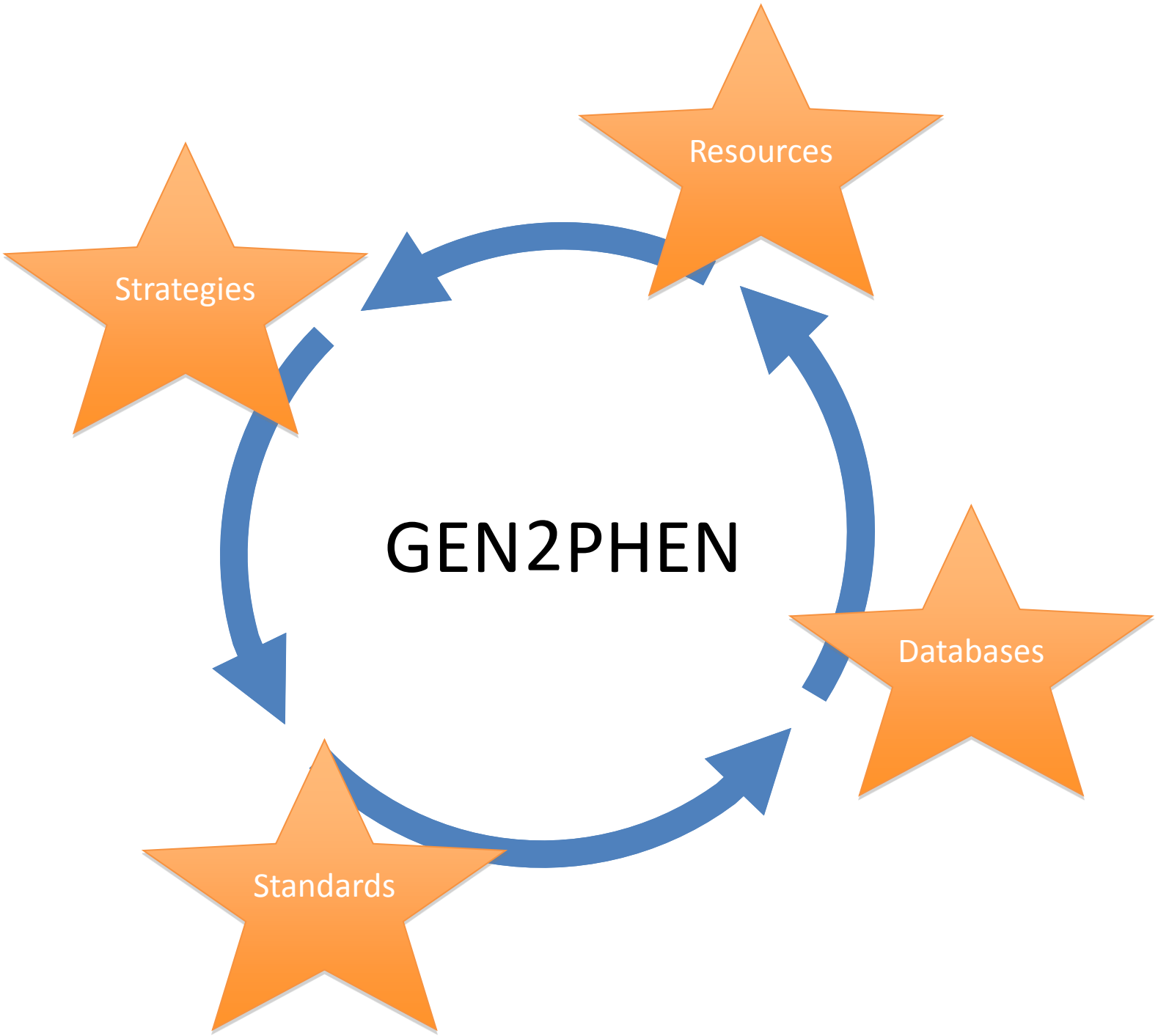
GEN2PHEN Partners (www.gen2phen.org)

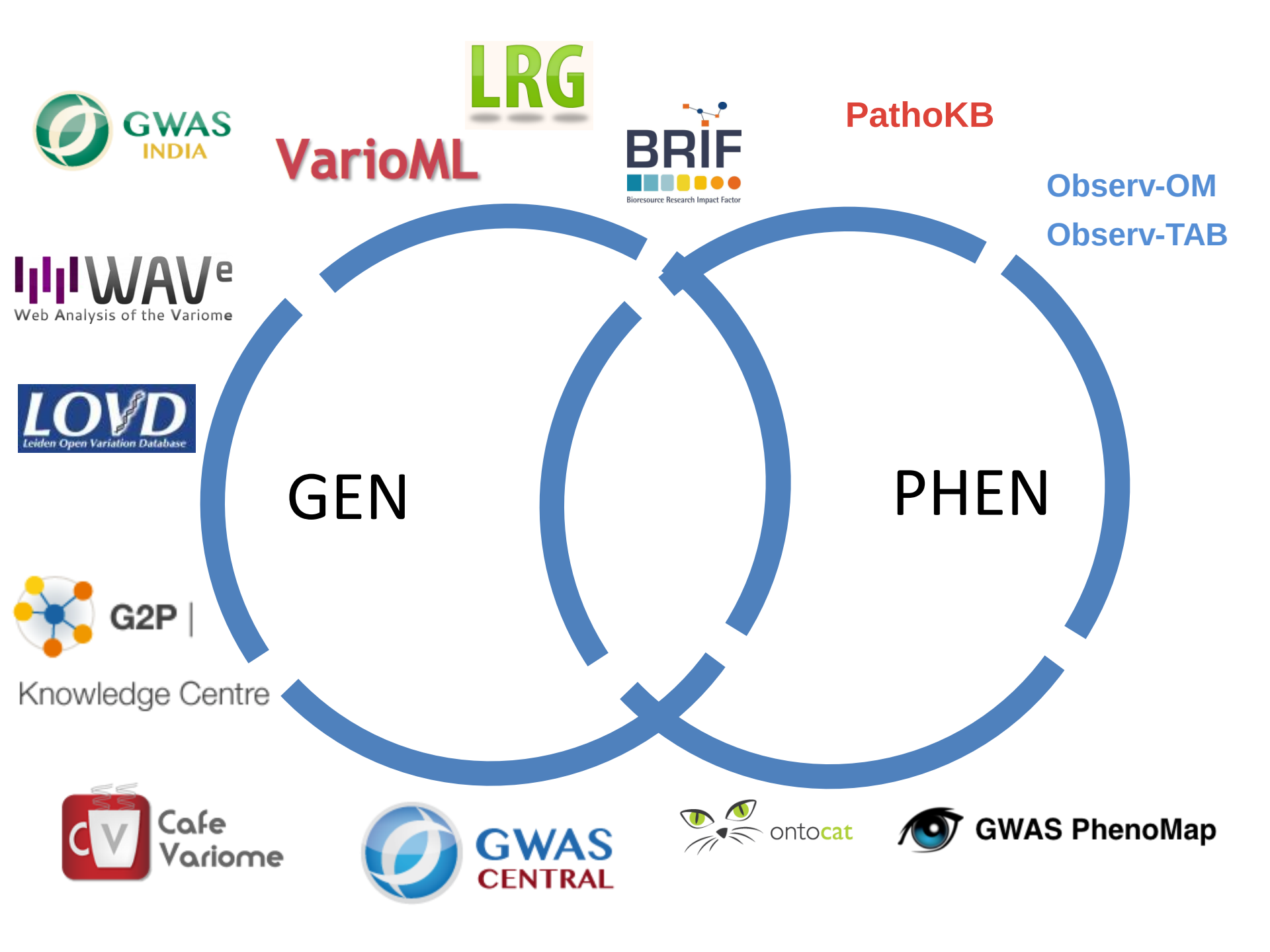
Academic		
A.J.Brookes, R.Dalgleish	University of Leicester	UK
P.Flicek, H.Parkinson	European Molecular Biology Laboratory	Germany
C.Díaz	Fundació IMIM	Spain
J.denDunnen		nds
C.Béroud		
A.Cambon-Thomsen		
J-E.Litton		
G.Potamias		
G.Patrinos		
M.Lathrop		
J.Muilu	University of Helsinki	Finland
J.L.Oliveira	University of Aveiro – IEETA	Portugal
D.Dash		
L.Yip		and
A.Devereau		
M.Swertz		nds
M.Vihinen		
SMEs		
A.Kel		y
H.Gudbjartsson	deCODE genetics	Iceland
D.Atlan	PhenoSystems	Belgium
T.Kanninen	Biocomputing Platforms	Finland
Previous		
H.Lehvaslaiho	University of Western Cape	South Africa

Towards a federated internet 'Knowledge-Environment' for genotype-phenotype data...

...for research and healthcare utilisation.


Resources
Strategies
GEN2PHEN
Databases
Standards


Resources
Strategies
GEN2PHEN
Databases
Standards


LRG
GWAS INDIA
VarioML
BRIF
Bioresource Research Impact Factor
PathoKB
Observ-OM
Observ-TAB
WAVe
Web Analysis of the Variome
LOVD
Leiden Open Variation Database
GEN
PHEN
G2P |
Knowledge Centre
Cafe Variome
GWAS CENTRAL
ontocat
GWAS PhenoMap

GEN2PHEN - Unification and utility of 'Genotype-to-Phenotype' data

BioShaRE - Science and IT for Biobanking

EMIF - Using health data in research

eTOX - Toxicology data and modelling

COPD-MAP - Systems medicine KM (COPD)

REQUITE - Systems medicine KM (cancer)

Research data sharing

Diagnostic lab data sharing

EHR data sharing

Pharma company data sharing

Public sourced data sharing

DATA SHARING IS IMPORTANT BUT DIFFICULT !

Nuffield Council on Bioethics workshop (12 Sept 2013)
"Fact finding meeting on privacy and data control issues"

- **General:**
  - IT can provide solutions, but political will to implement complicated by rapid changes in IT
  - nothing can offer one hundred per cent protection
  - attitudes to data protection vary considerably across different cultures

- **Data Warehouses:**
  - centralisation of data has made it easier to illicitly access private information
  - frequent data security breaches creates rational public distrust in many data-heavy projects
  - there is a significant black market for personal data

- **Patients:**
  - do not necessarily value privacy to the exclusion of all other interests
  - it is not clear how well they understand the issues
  - are more likely to offer data if given something in return

- **Researchers (carrots, sticks, ambition, ability):**
  - ????????

- **Organisations (business model and legal considerations):**
  - ????????

Data Sharing… just needs bett gineering?

metadata standards

federation standard

exchange f

in

use cat

ethico-social

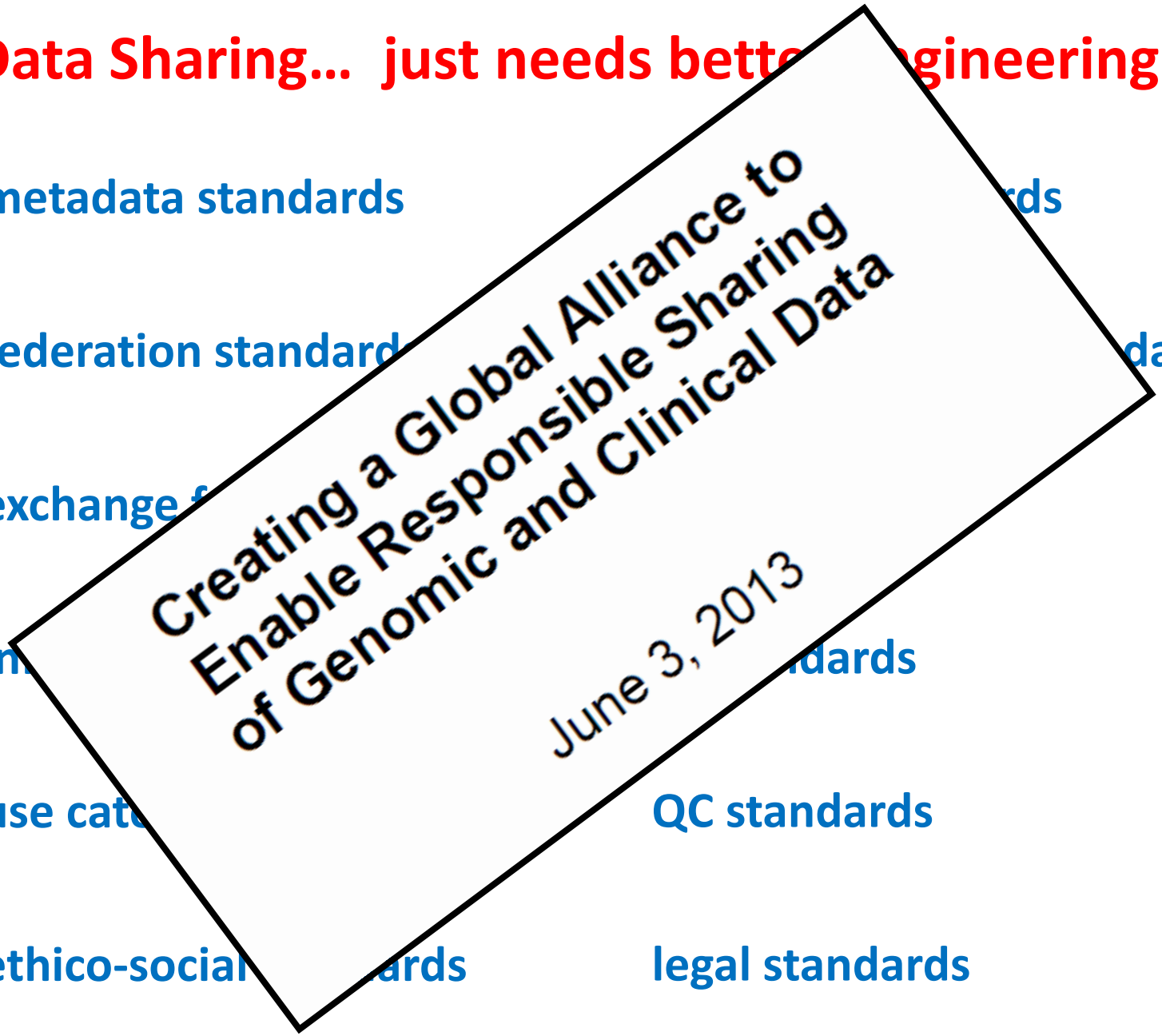
rds

dards

dards

QC standards

legal standards



ENTITY IDENTIFIERS

Data IDs

- The 'Data Object Identifier' (DOI) system, managed by DataCite. Covers a very broad concept of a 'data object' (much more than just traditional publications). Essential for creating the 'web of data'.

Database IDs

- The BioDBCore project by which database IDs can be assigned. Essential if webservices are to start connecting resources effectively.

Human IDs

- The 'Open Researcher Contributor Identifier' (ORCID) system. Launched late 2013, has already issued many tens of thousands of ORCIDs. Soon to be a required author detail when submitting manuscripts. Removes ambiguity over all 'contributors', thereby enabling incentive/reward systems for data sharing, improved knowledge discovery options, and automation of data access control.

Patient IDs

- Various approaches being tested, but none yet close to an ideal, global solution. Needed to resolve identical patient records that relate to different individuals from those that are merely duplicated (recycled) information about a single individual.

BioResource IDs

- Pilot system emerging from GEN2PHEN & BioShaRE, operated by P3G, as a basis for developing BioResource Impact Factor (BRIF) metrics.

Issues that restrict sharing data

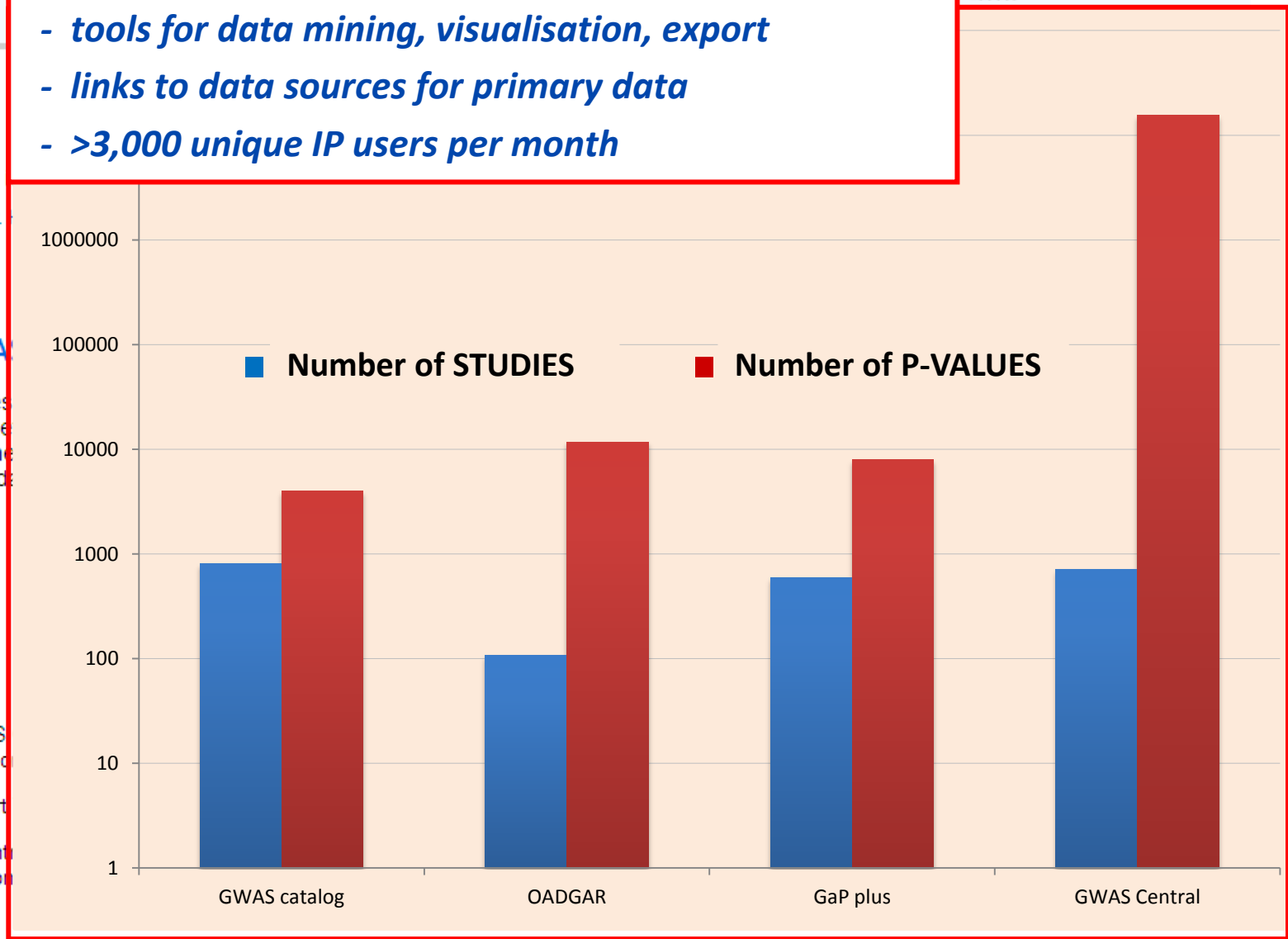
- CANNOT: Data owners may not have time nor funding to manually submit data, and/or submission process and requirements too complicated

- WILL NOT: Data owners receive little or no recognition or reward for releasing data, hence little incentive to try

- MUST NOT: Data owners may have good reasons for not sharing data (ethical, legal, competitive edge)

‘GWAS Central’

- *GWAS db: 1700 studies + 30,000,000 p-values*
- *aggregate data & extensive metadata*
- *tools for data mining, visualisation, export*
- *links to data sources for primary data*
- *>3,000 unique IP users per month*



GWAS CENTRAL 'Publicity Project'

March 2011

570 e-mails sent to authors of 700 studies [30 undeliverable]

19 (3.5%) positive responses (new or corrected data)

9 (1.7%) neutral or negative responses

512 (94.8%) did not reply

January 2012

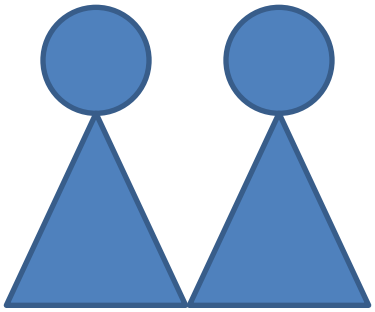
505 e-mails sent to authors of 512 studies [16 undeliverable]

17 (3.5%) positive responses (new or corrected data)

38 (7.8%) neutral or negative responses

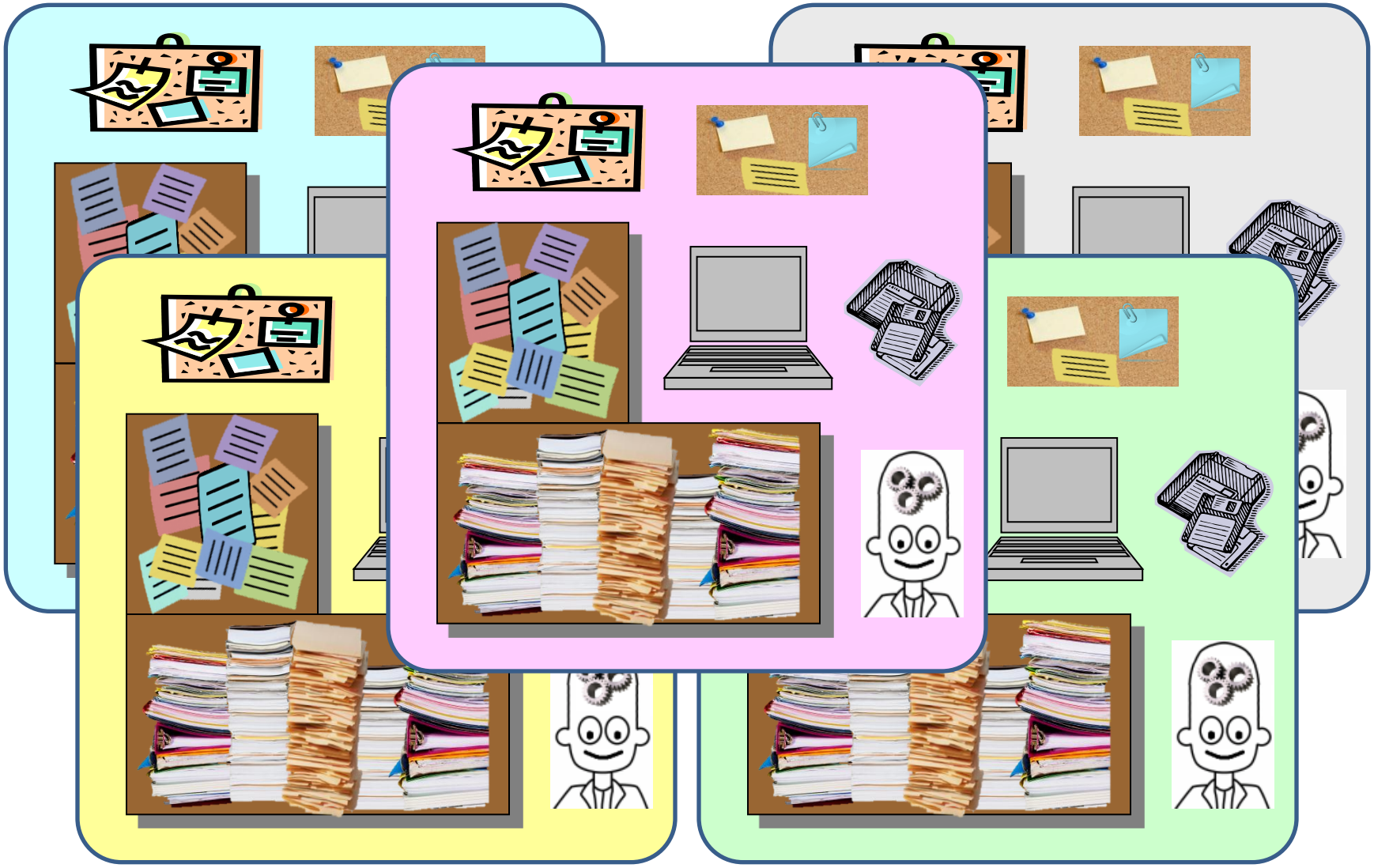
434 (88.7%) did not reply

Improving on the current state-of-the-art



Single
Research Team

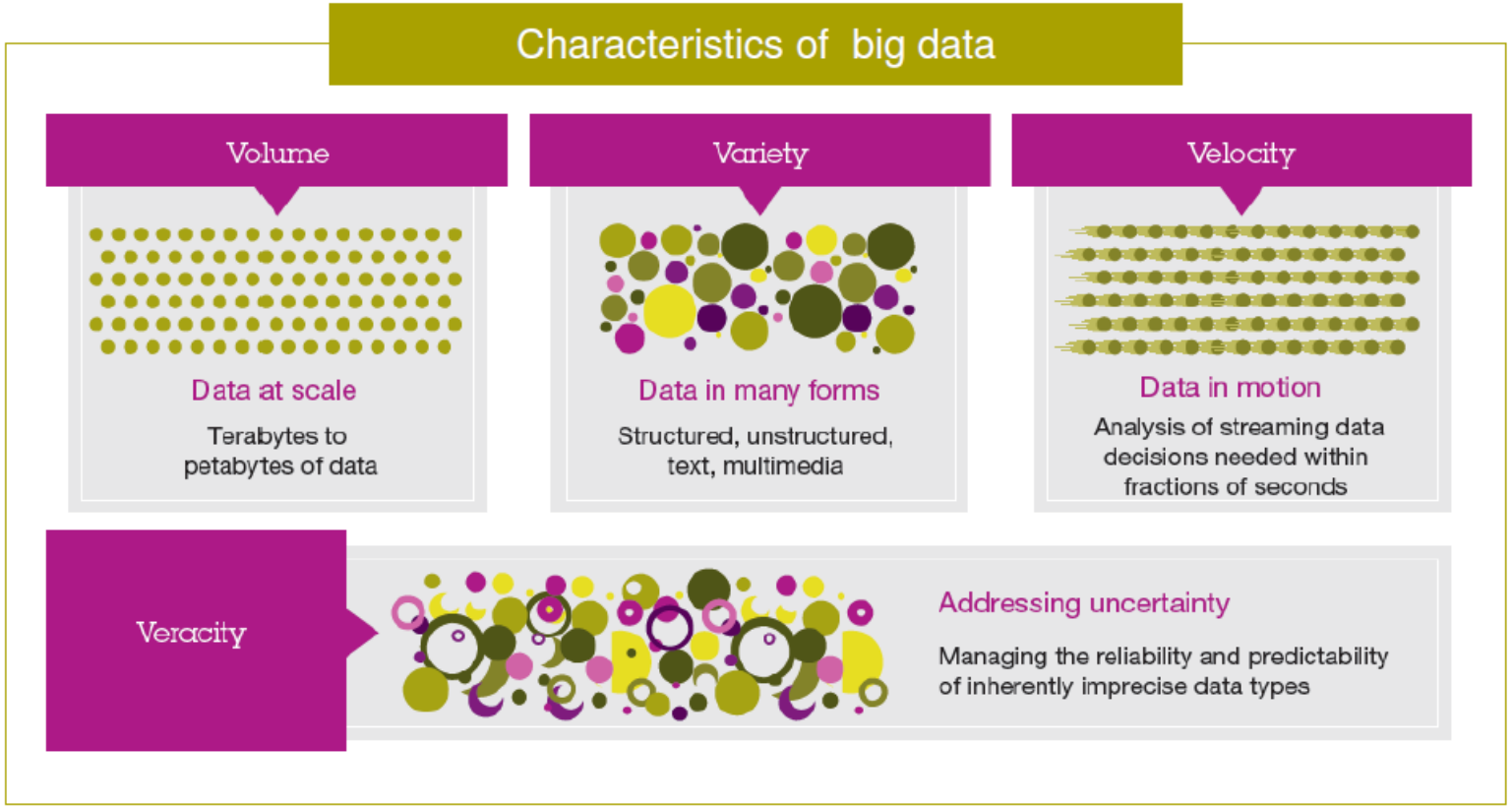
Data



Consortium Data

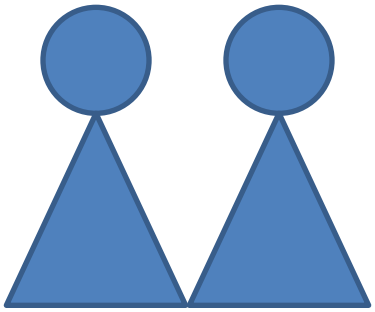

‘Big Data’

5. BIG DATA...



 |

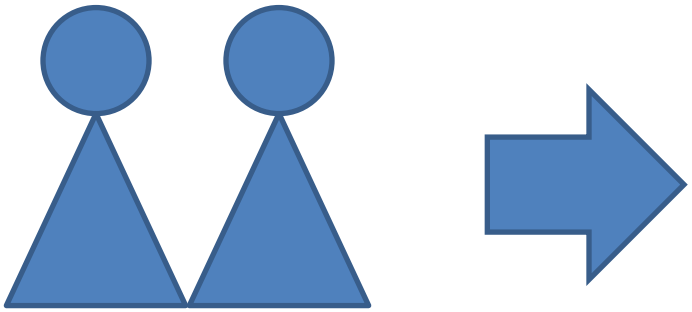
IBM



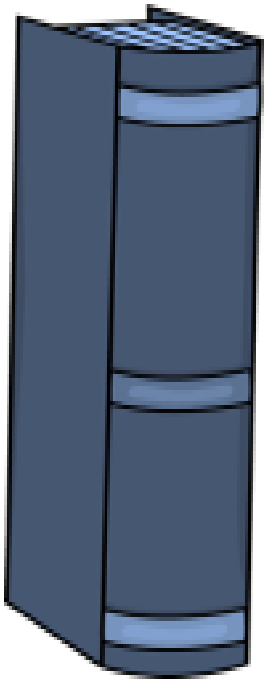
Single
Research Team

Data

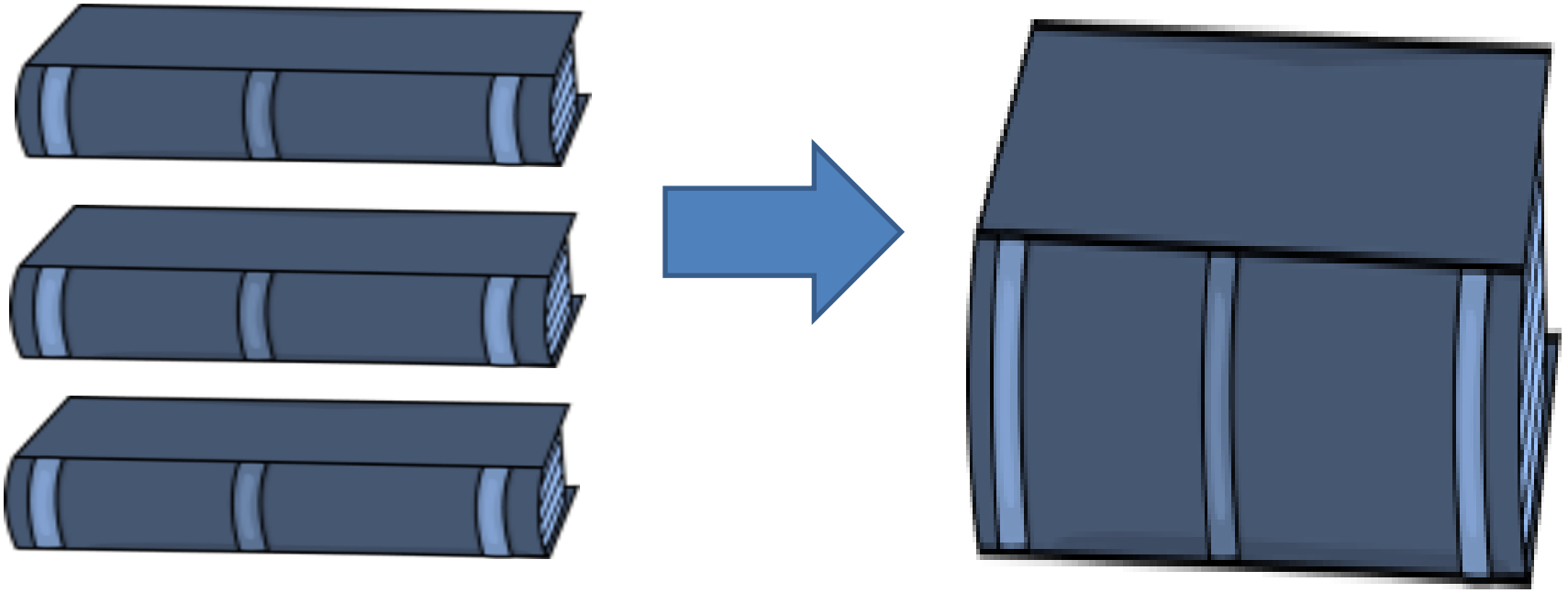
Can/Will all data be digitised, organised, structured... to make them 'sharable' ?



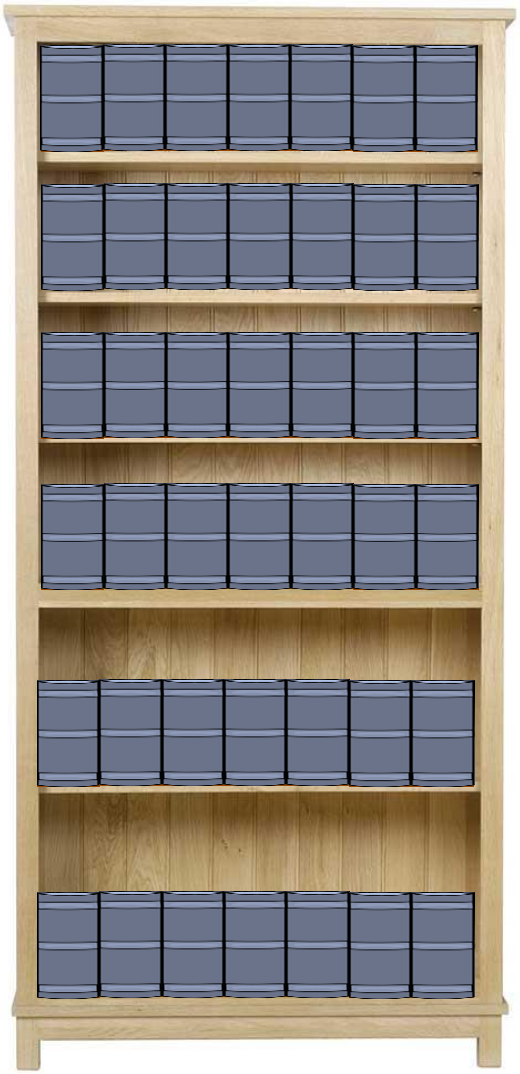
Research Team



Data



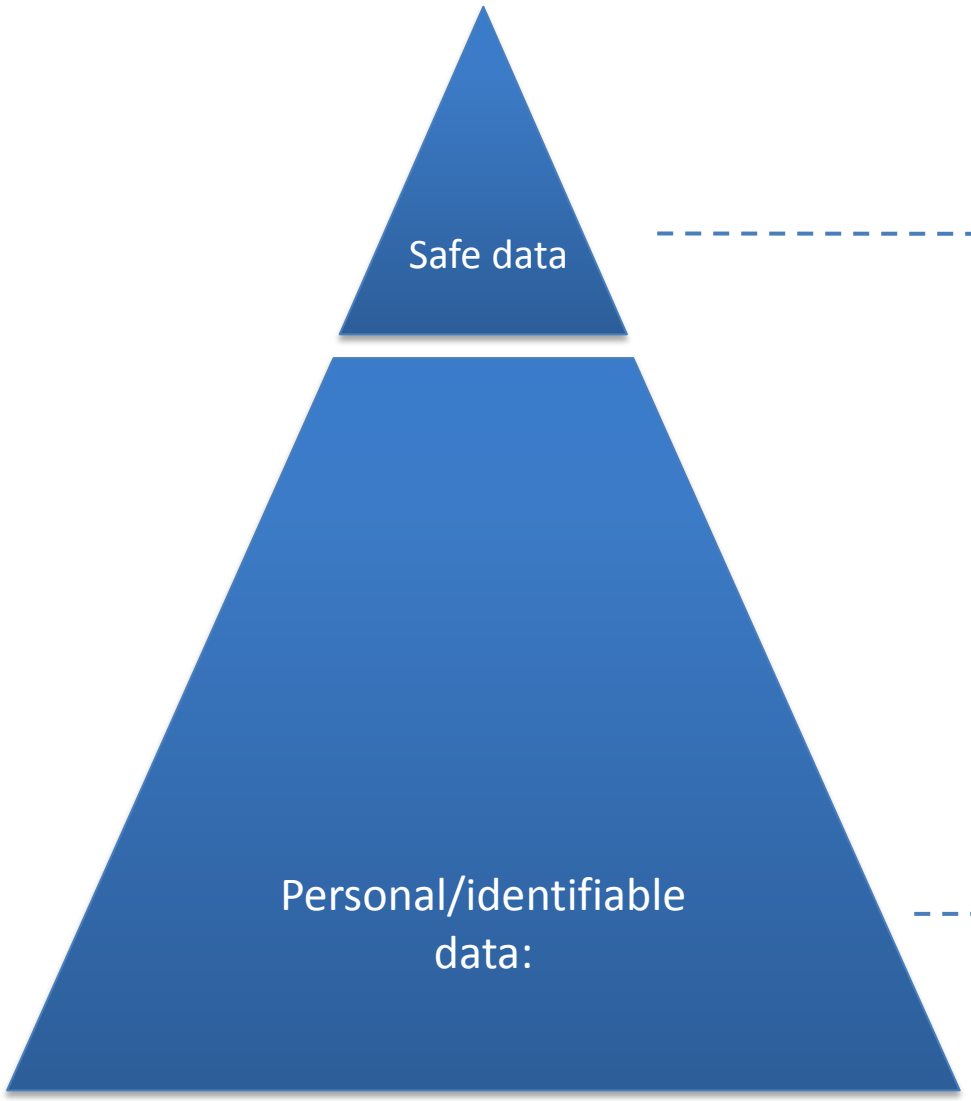
Centralised Sharing of Data



Federated Sharing of Data

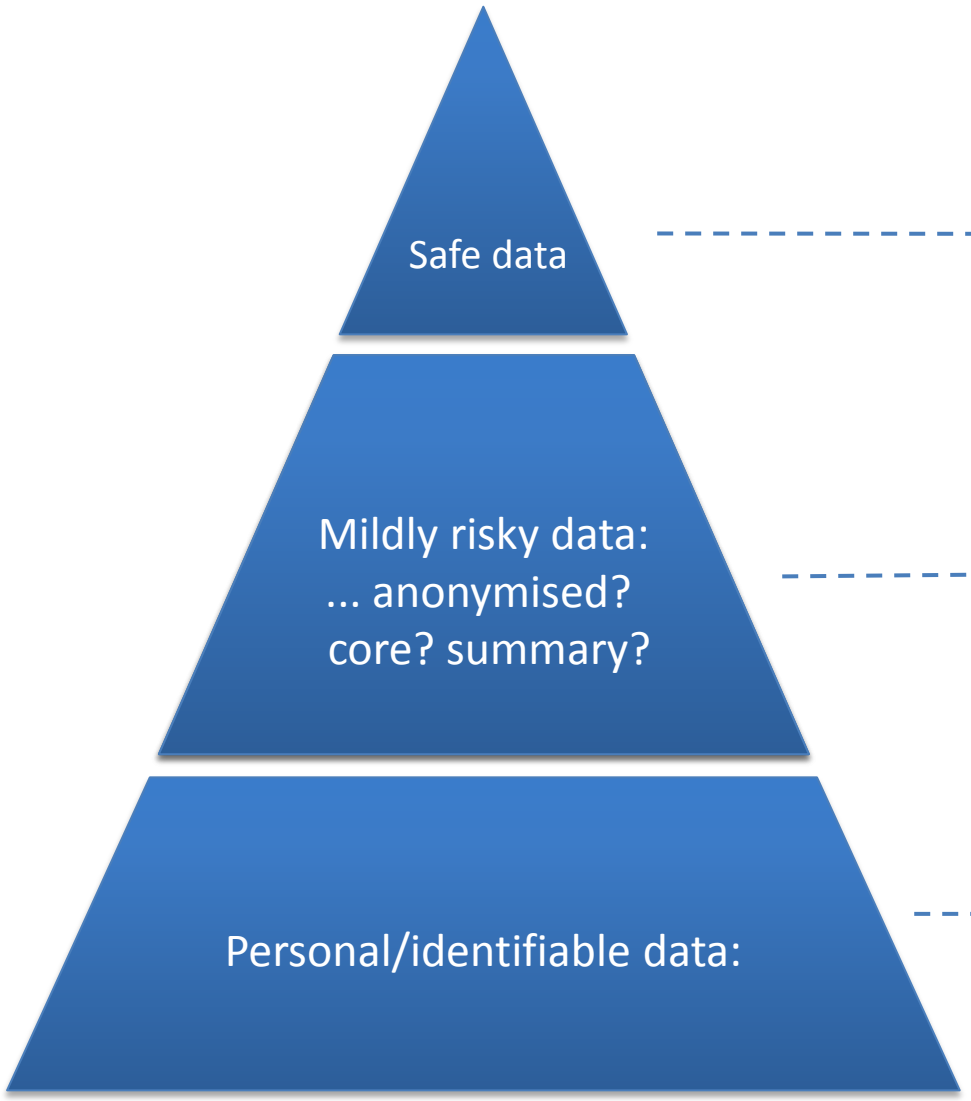
Realistic Expectation

SHARING
(of data)



open sharing

managed access (EGA, dbGaP)



	SHARING (of data)	DISCOVERY (of data & knowledge)
Safe data	open sharing	open
Mildly risky data: ... anonymised? core? summary?	Open or ‘speed pass’ (ID assisted, automated)	open
Personal/identifiable data:	managed access (EGA, dbGaP)	open

Openly share the 'existence' rather than the 'substance' of the data

….thereafter variably manage data access

Connecting Diagnostic Networks

- Need to enable diagnostic laboratories to check whether sequence variants have previously been seen by other labs [with patients with related phenotypes]

- Currently not possible due to difficulties of data sharing between labs, or with others

- Café Variome can solve this...

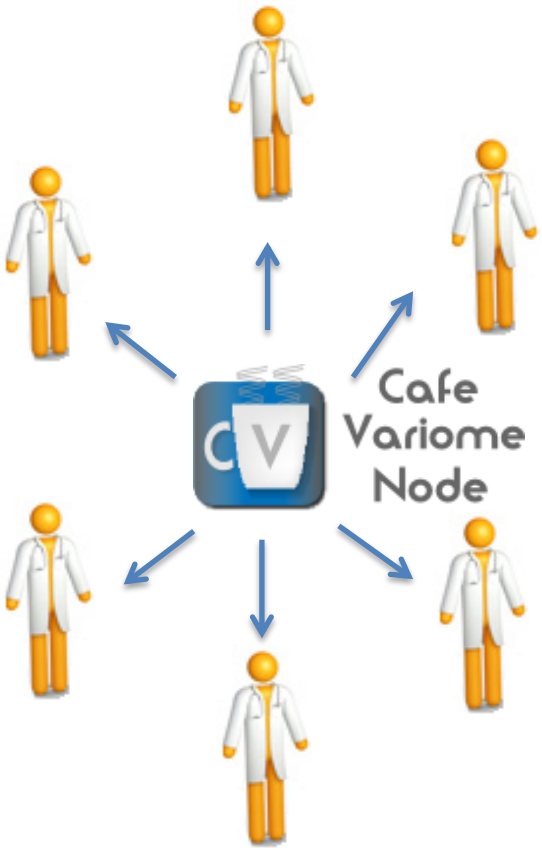
The Café Variome Solution

- Allows 'open discovery' of the existence (rather than actual substance) of relevant mutation data

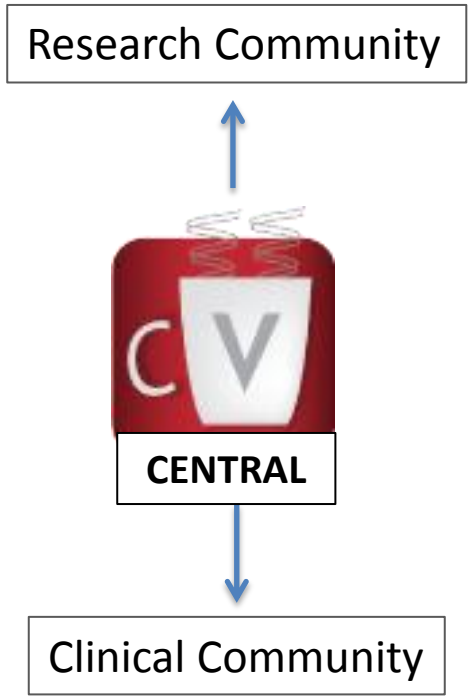
- Thereby, enables networks of labs to easily query for the existence of variants, without necessarily revealing the underlying data, thus overcoming issues of patient confidentiality, etc.

- Currently being extended to handle more sophisticated omics/NGS data handling and deep phenotype data

Networks of labs exchanging data



Optional



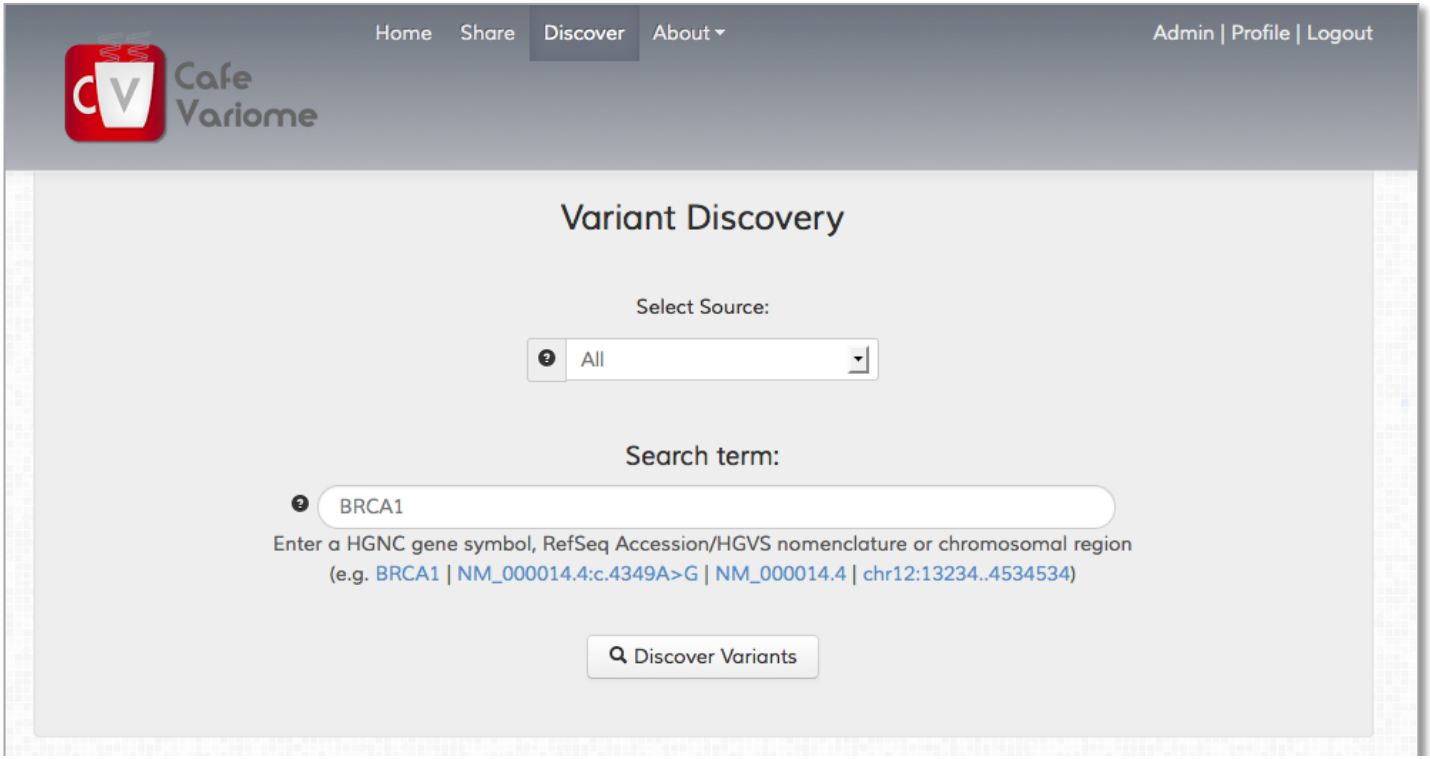
Allows users to check whether the same mutation(s) have previously been seen by other laboratories/patients (with related phenotypes)

Café Variome Central (http://www.cafevariome.org)

Source	Variant Count
Diagnostic laboratories	193
HGMD	57927
UniProt Variants	23558
1000 Genomes	86305
dbSNP (coding)	100103
PhenCode	3694
DMuDB (pilot data)	21
LSDB	57868
FORGE Canada	73
FINDbase	6850
TOTAL:	**336592**

Café Variome Central is now populated with all major pubic human genome sequence variants, including all public LSDBs

Record Discovery “Menu”

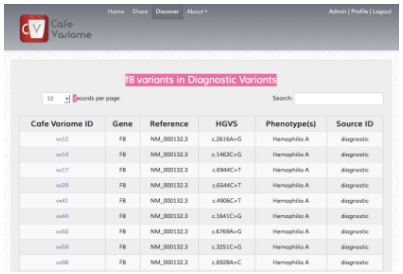


Data Sharing Models (facilitated, controlled access)

Open Discovery – Reporting Existence of Variants in Sources

Open Access

Core info for each variant record is shown & made available for download



Restricted Access

Core or full record details are provided per record, if:

- User is pre-approved by group access permissions set by data owner
- Access is approved after facilitated email request to the data owner

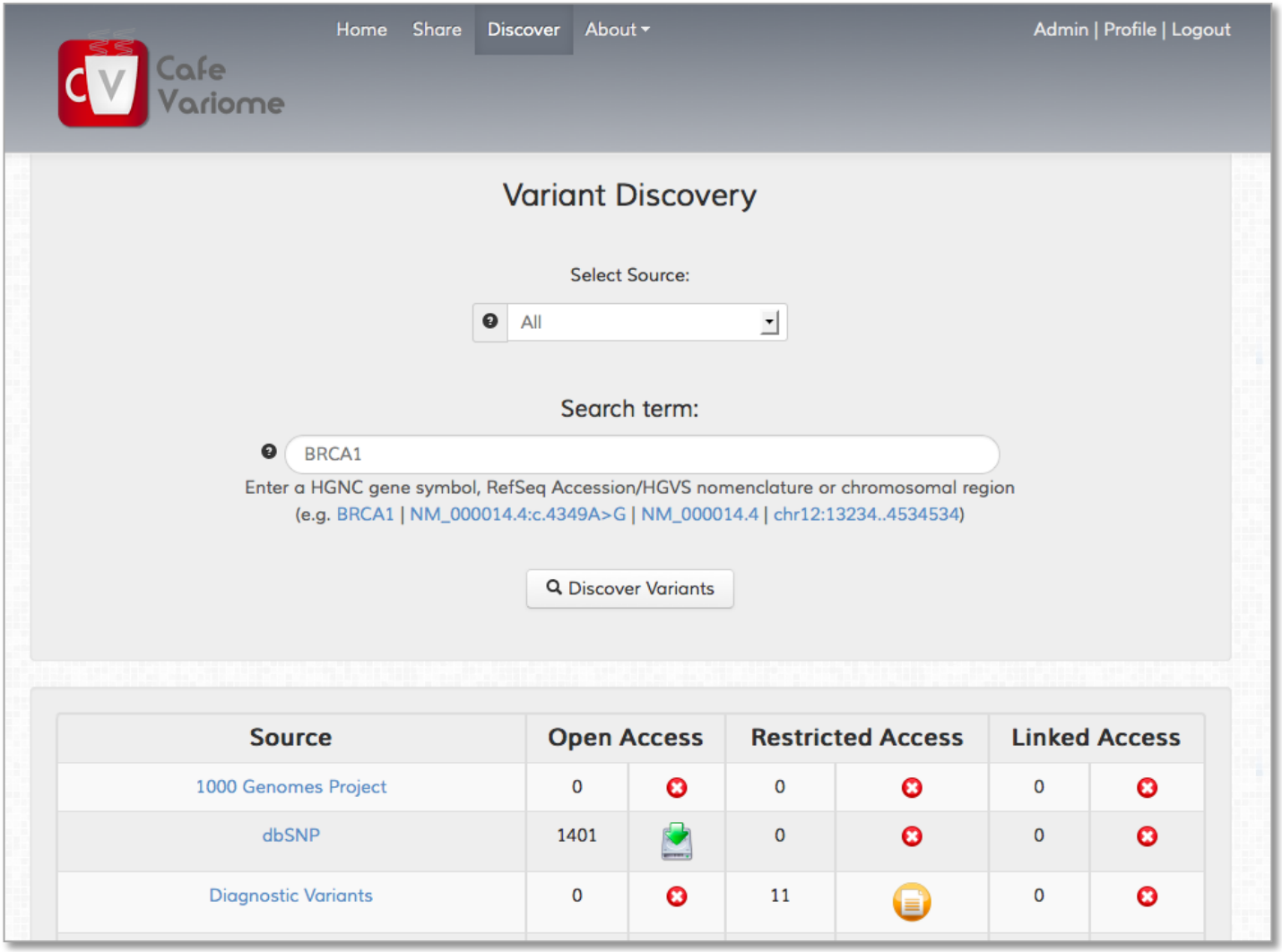
Linked Access

Source DB resource

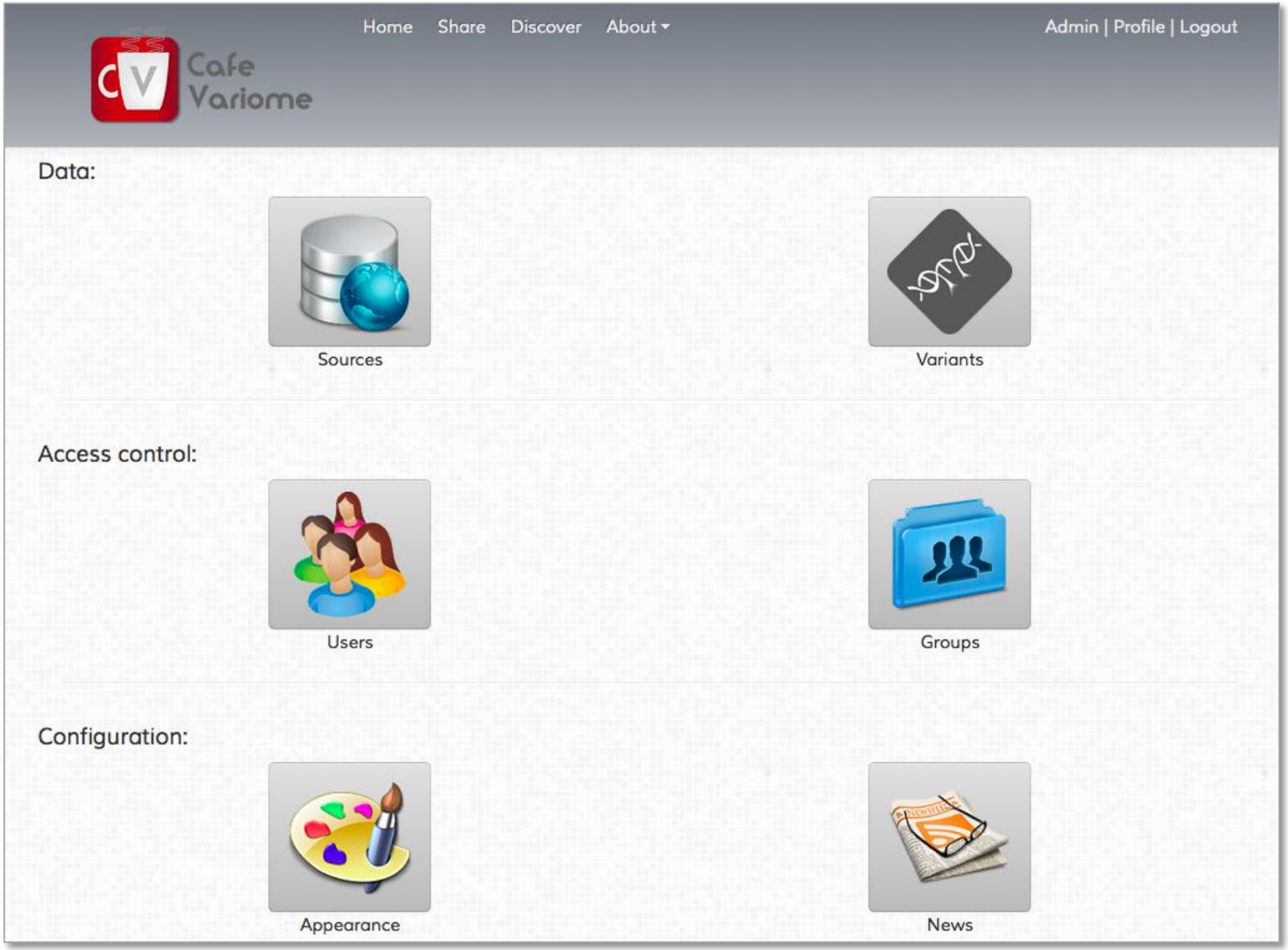
No data, only link to the data source is reported.

Access then control managed by source db

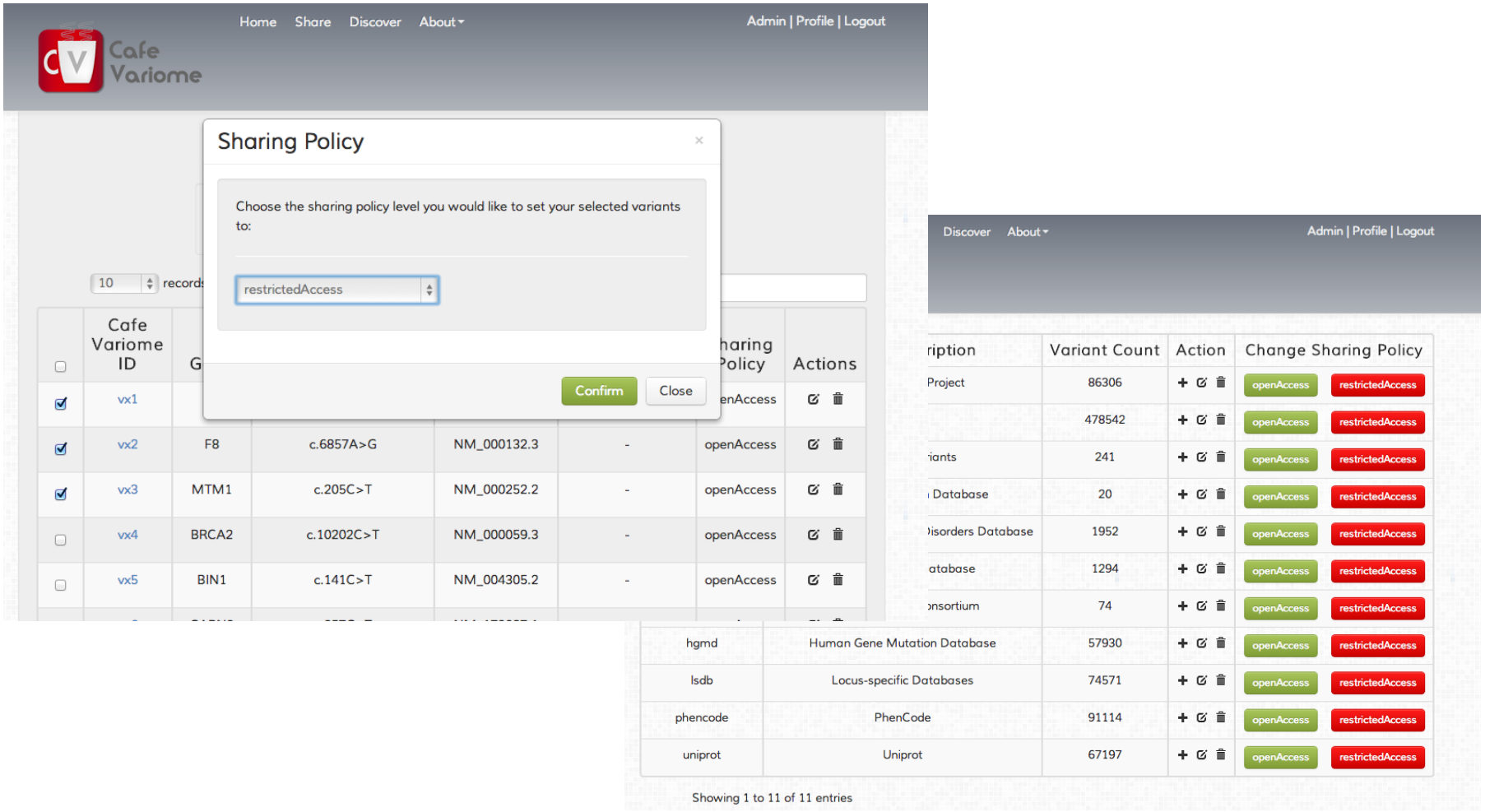
Record Discovery “Menu”



Administrators Interface



Data Sharing Granularity



Users can control access to variants from individual fine grained level to whole data sets

Café Variome adoption and use

- **Requires only three input fields per recorded mutation**
  - HGVS name, gene name & reference URL
  - optional phenotype, patient ID, & other fields
  [metadata includes source DB name, curator name, reference transcript]

- **Directly installed from a simply installation file**

- **Supports networks by being installed**

  - on a single server
    - multi-site reference data loaded into this menu
    - multi-site reference data kept locally & accessed remotely by web-services

  - on multiple servers
    - multi-site reference data kept locally & searched from these menus

- **Data also loaded directly by upload buttons available on**
  - Alamut software (Interactive Biosystems)
  - Gensearch (PhenoSystems)

Cafe Variome Network Collaborators

- DMuDB (with Andrew Devereau, Susan Stenhouse)
- Denmark diagnostic network (with Friedrik Wikman)
- German diagnostic network (with Arne Pfeufer)
- French diagnostic network (with Andre Blavier (Interactive Biosoftware))
- Belgium diagnostic network (with Gert Matthijs)
- Canadian diagnostic network (with Forge Canada / Care4Rare)
- Netherlands diagnostic network (with Morris Swertz & Rolf Sijmons)

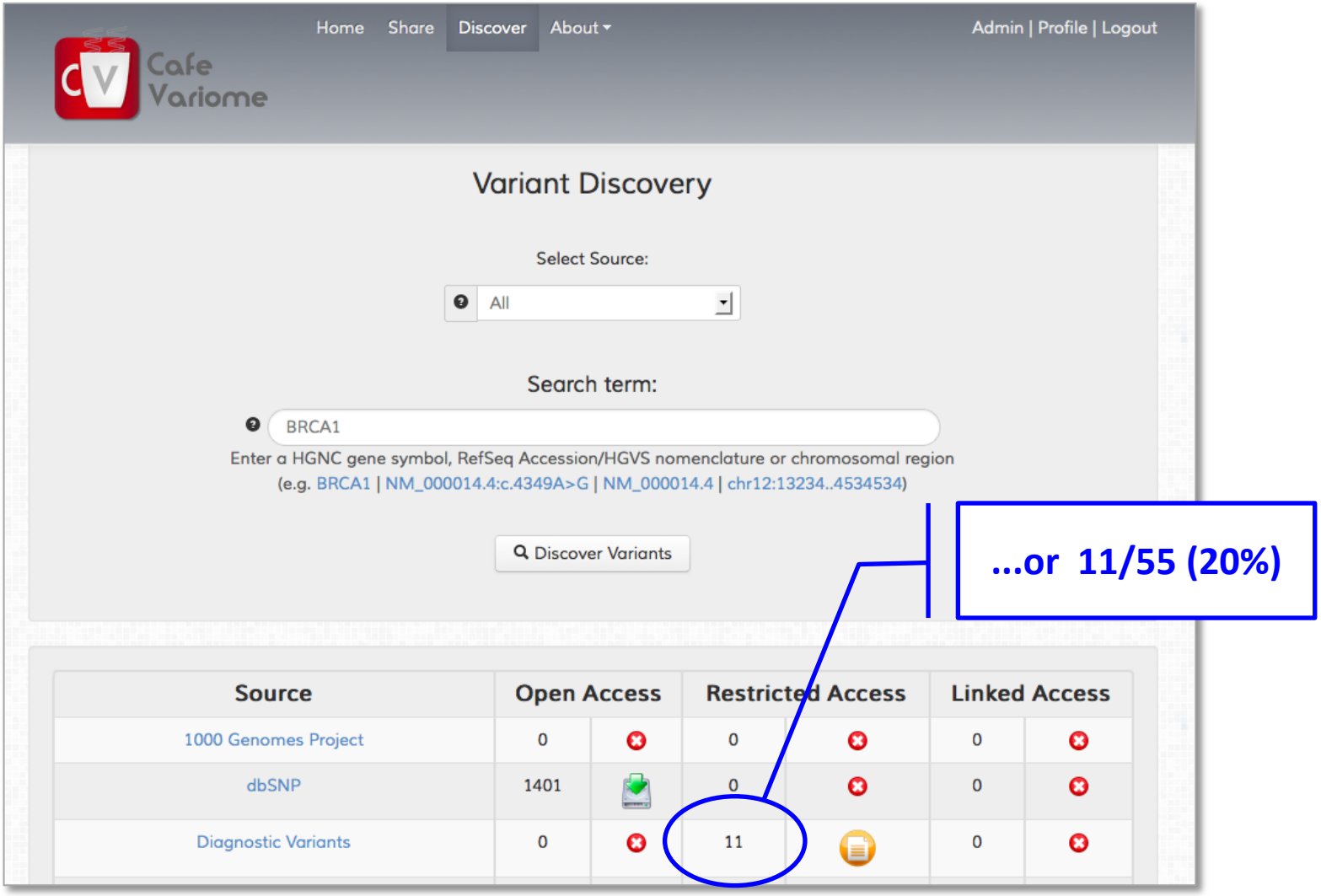
- Ehlers Danlos Syndrome network (with Raymond Dalgleish)
- Collagen disease network (with Peter Byers)
- Inherited Colon Cancer network (with Finlay Macrae, InSiGHT)

Data Discovery

-->

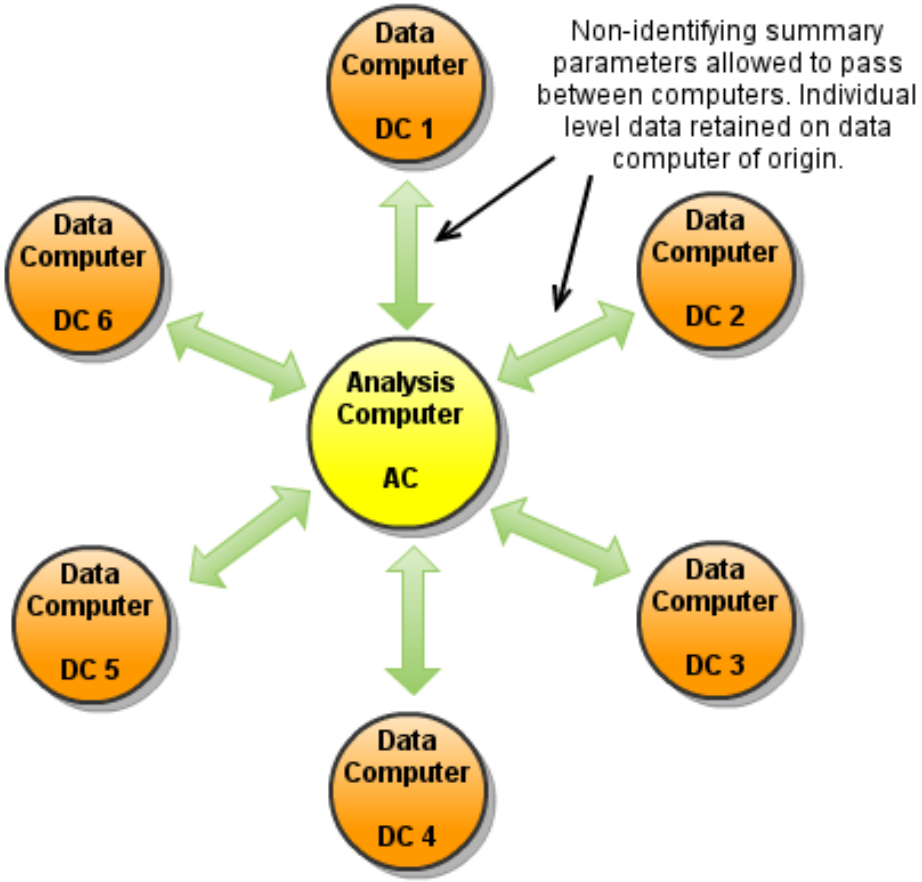
Knowledge Sharing

Record Discovery “Menu”

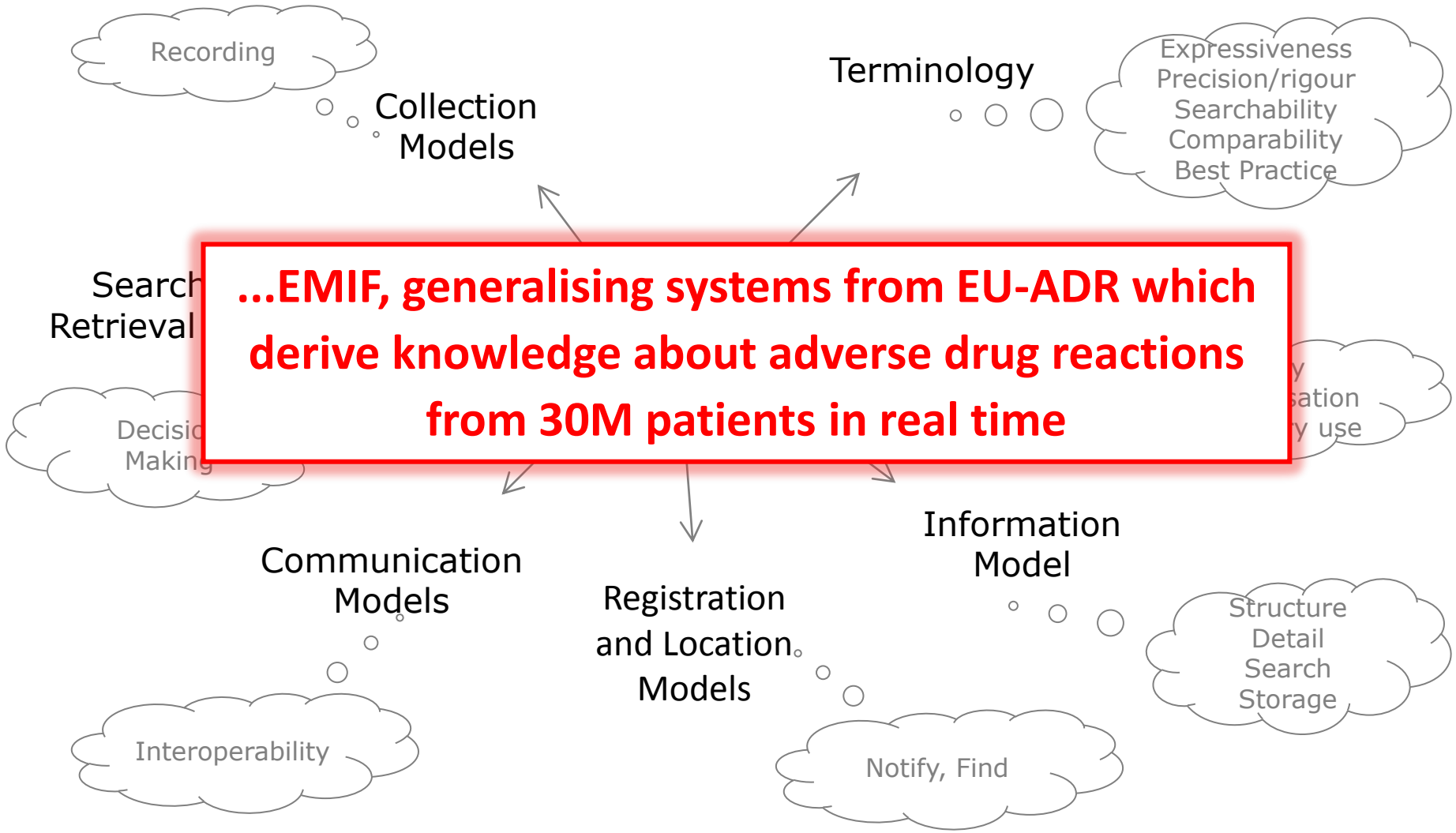


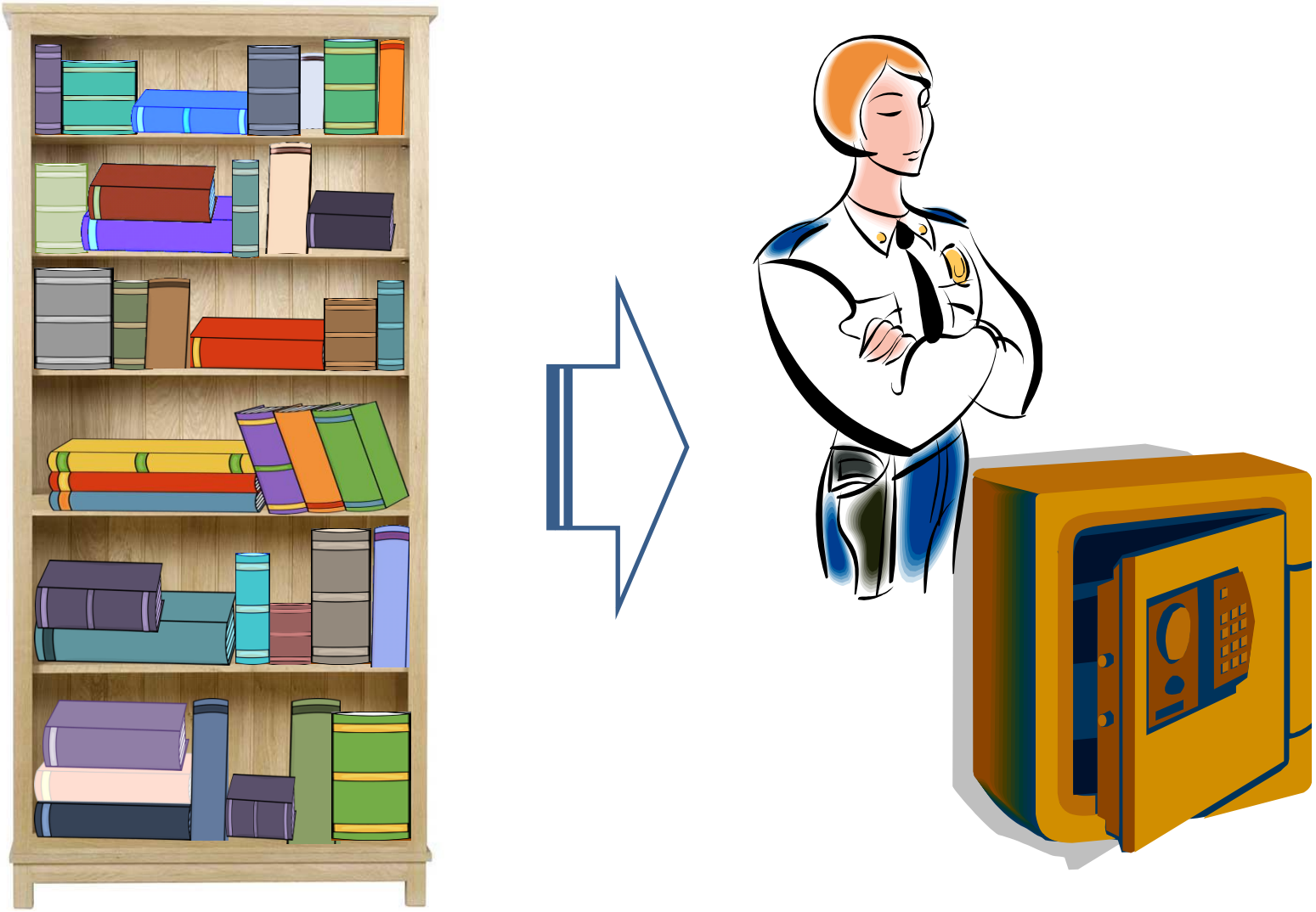
DataSHIELD: Pooled data analysis without data sharing

- An Analysis Computer (AC) sends iteratively requests for fitting a given model to the Data Computers (DC) on which data are stored



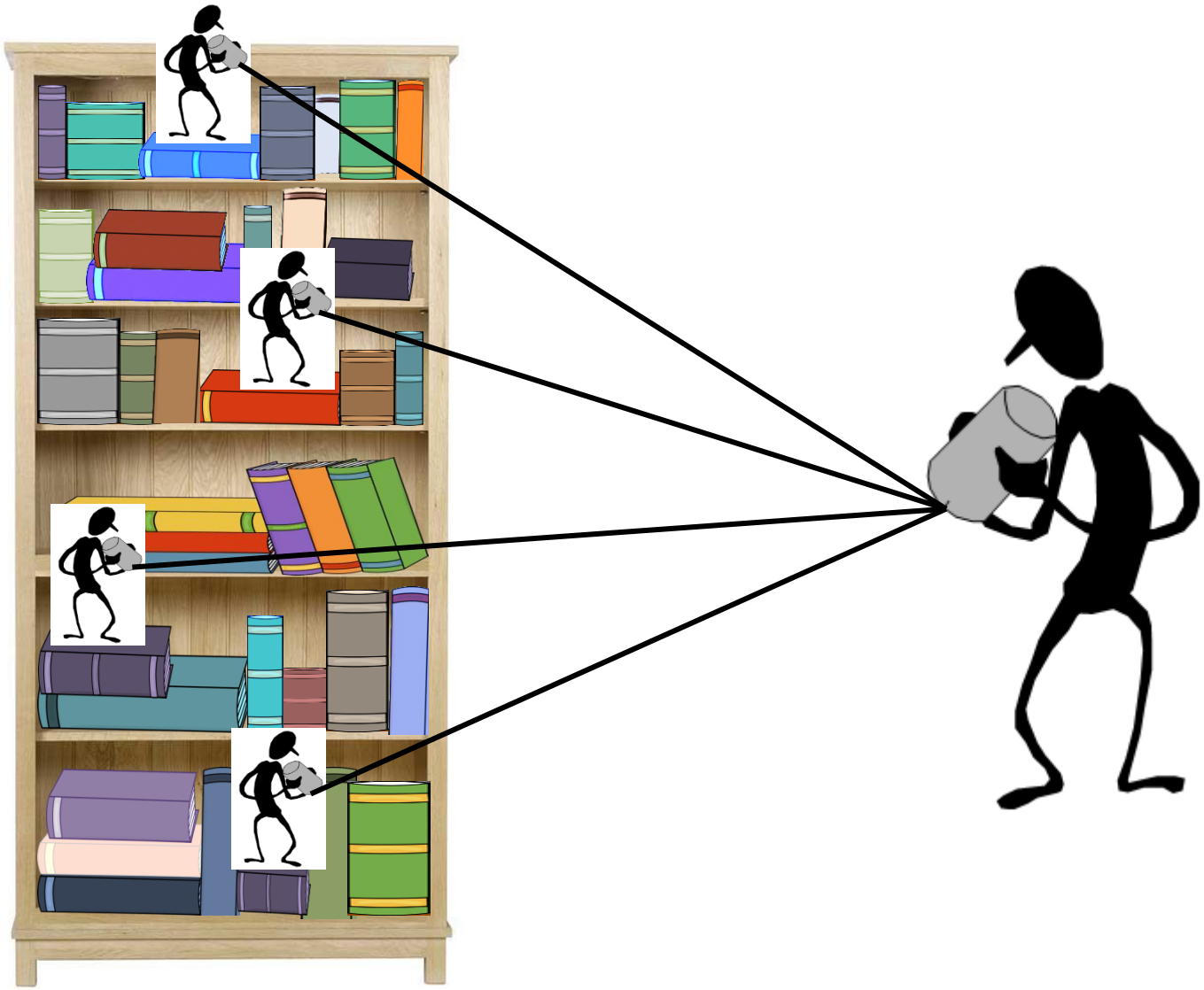
Knowledge Sharing



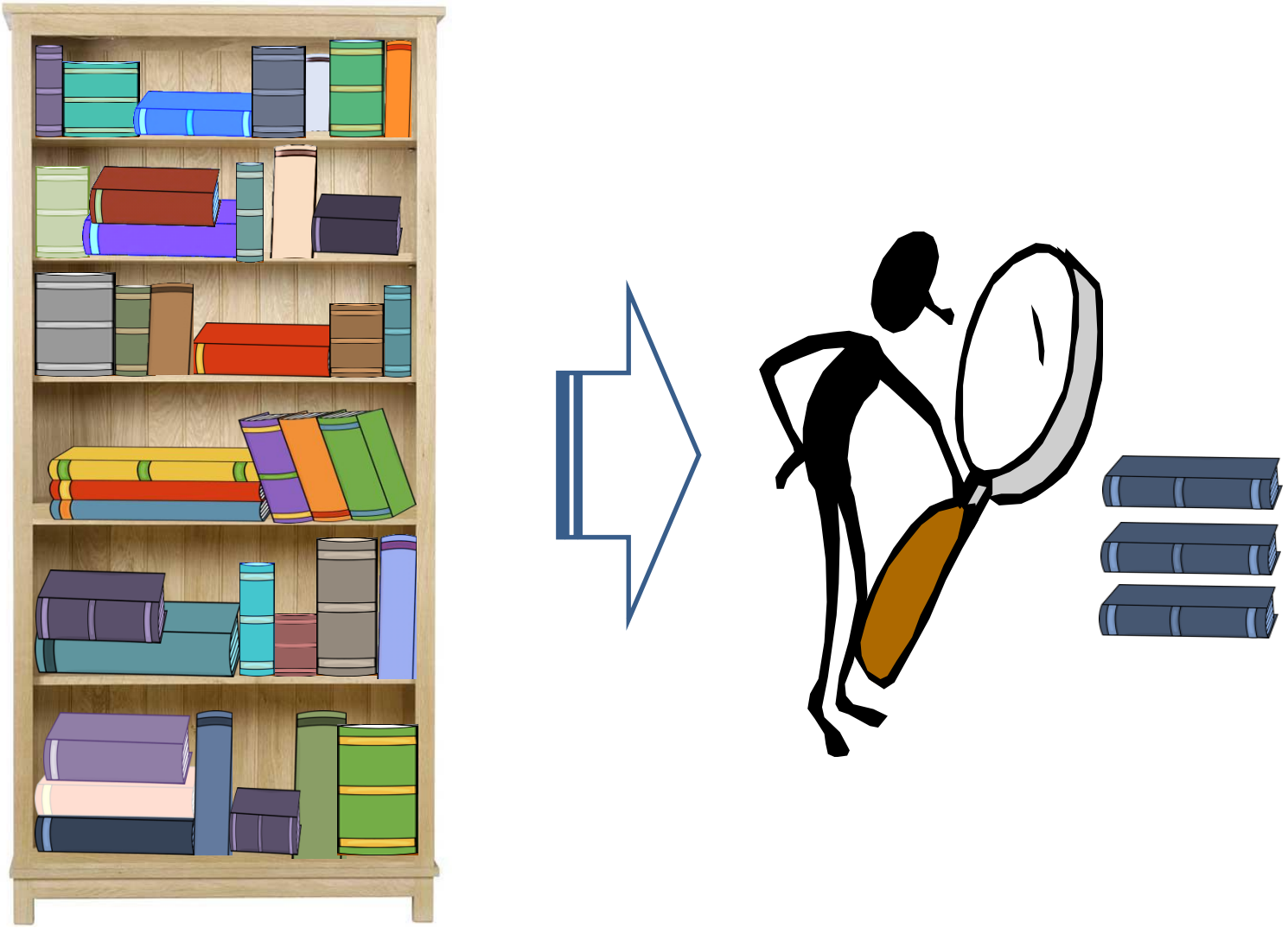


Data Sharing

Knowledge Sharing



Data Remote Analysis



Data Discovery

DATA - important, but problematic
SHARING - prioritise IDs and risk categorisation

DATA - 90% of the benefits
DISCOVERY - 10% of the problems

KNOWLEDGE - just a technical extension
SHARING - 99% of the benefits

One does not have to co-locate:
- Data repository & Data interpretation

One does not have to co-locate:
- Data sharing & Knowledge sharing & Data discovery

University of Leicester

Partners in GEN2PHEN, BioShaRE, EMIF, eTox, COPD-MAP, REQUITE

Anthony Brookes' Bioinformatics Group

Tim Beck, Charalambos Chrysostomou, Robert Hastings, Owen Lancaster, Adam Webb

Sirisha Gollapudi, Robert Free, Gudmundur Thorisson

DKP Facility members

Data to Knowledge for Practice (DKP) Facility